UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

1992

25ème anniversaire

N° 1814

Programme 2
Calcul Symbolique, Programmation
et Génie logiciel

A PROBABILISTIC ANALYSIS OF A STRING EDIT PROBLEM

Guy LOUCHARD
Wojciech SZPANKOWSKI

Décembre 1992

A PROBABILISTIC ANALYSIS OF A STRING EDIT PROBLEM

Guy Louchard
Laboratoire d'Informatique Théorique
Université Libre de Bruxelles
B-1050 Brussels
Belgium

Wojciech Szpankowski*
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Abstract

We consider a string edit problem in a probabilistic framework. This problem is of considerable interest to many facets of science, most notably molecular biology and computer science. A string editing transformes one string into another by performing a series of weighted edit operations of overall maximum (minimum) cost. An edit operation can be the deletion of a symbol, the insertion of a symbol or the substitution of a symbol. We assume that these weights can be arbitrary distributed. We reduce the problem to finding an optimal path in a weighted grid graph, and provide several results regarding a typical behavior of such a path. In particular, we observe that the optimal path (i.e., edit distance) is asymptotically almost surely (a.s.) equal to αn where α is a constant and n is the sum of lengths of both strings. We also obtain explicit bounds on the constant α. More importantly, we show that the edit distance is well concentrated around its average value. As a by-product of our results, we also present a precise estimate of the number of alignments between two strings. To prove these findings we use techniques of random walks, diffusion limiting processes, generating functions, and the method of bounded difference.

ANALYSE PROBABILISTE DU PROBLEME DE L'EDITION DE MOTIF

Résumé

Nous considérons le probléme de l'édition d'un motif (une chaîne de caractéres) dans un environnement probabiliste. Ce probléme est d'un grand intérêt dans beaucoup de domaines scientifiques principalement en biologie moléculaire et en informatique. Une édition de motif transforme un motif en un autre en réalisant une suite d'opérations d'édition, pondérées, de coût total maximum (ou minimum). Une telle opération peut être: l'élimination d'un symbole, l'insertion ou la substitution d'un symbole. Nous supposons que les poids peuvent être distribués aléatoirement. Nous réduisons le probléme à celui de trouver un chemin optimal dans un treillis pondéré, et nous obtenons plusieurs résultats relatifs au comportement typique d'un tel chemin. En particulier, nous observons que le chemin optimal (c'est-à-dire la distance d'édition) est asymptotiquement, presque sûrement, égal à αn où α est constant et n est somme des longueurs des deux motifs. Nous obtenons également des bornes sur α. Plus important, nous montrons que la distance d'édition est trés concentré autour de sa moyenne. En corollaire de nos résultats, nous présentons également un estimateur précis du nombre d'alignement entre les deux motifs. Pour démontrer nos résultats, nous utilisons des techniques de chemin aléatoire, de processus limites de diffusion, de fonctions génératrices et la méthode des différences bornées.

*Support for this research was provided in part by by NSF Grants CCR-9201078, NCR-9206315 and INT-8912631, in part by AFOSR Grant 90-0107, NATO Grant 0057/89, and Grant R01 LM05118 from the National Library of Medicine. This research was completed while the second author was visiting INRIA, Rocquencourt, France, and he wishes to thank INRIA (projects ALGO, MEVAL and REFLECS) for a generous support.

1. INTRODUCTION

String editing problem arises in many applications, notably in text editing, speech recognition, machine vision and, last but not least, molecular sequence comparison. Algorithmic aspect of this problem has been studied rather extensively in the past (cf. [2], [28], [30], [31] and [34]). In fact, many important problems on words are special cases of string editing, including the *longest common subsequence* problem (cf. [1], [14]) and the problem of *approximate pattern matching* (cf. [12] and [32]).

In sequel we review the string editing problem, its importance, and its relationship to the longest path problem in a special grid graph.

Let **b** be a string consisting of ℓ symbols on some alphabet Σ of size V. There are three operations that can be performed on a string, namely *deletion* of a symbol, *insertion* of a symbol, and *substitution* of one symbol for another symbol in Σ. With each operation is associated a *weight* function. We denote by $W_I(b_i)$, $W_D(b_i)$ and $W_Q(a_i, b_j)$ the weight of insertion and deletion of the symbol $b_i \in \Sigma$, and substitution of a_i by $b_j \in \Sigma$, respectively. An *edit script* on **b** is any sequence ω of edit operations, and the total weight of ω is the sum of weights of the edit operations.

The *string editing problem* deals with two strings, say **b** of length ℓ (for ℓong) and **a** of length s (for $s$hort), and consists of finding an edit script ω_{max} (ω_{min}) of minimum (maximum) total weight that transforms **a** into **b**. The maximum (minimum) weight is called the *edit distance from* **a** *to* **b**, and its is also known as the Levenshtein distance. In molecular biology, the Levenshtein distance is used to measure similarity (homogeneity) of two molecular sequences, say DNA sequences (cf. [31]).

The string edit problem can be solved by the standard dynamic programming method. Let $C_{\max}(i,j)$ denote the maximum weight of transforming the prefix of **b** of size i into the prefix of **a** of size j. Then, (cf. [2], [28], [34]).

$$C_{\max}(i,j) = \max\{C_{\max}(i-1,j-1) + W_Q(a_i,b_j) \quad , \quad C_{\max}(i-1,j) + W_D(a_i) \ , \\ , \quad C_{\max}(i,j-1) + W_I(b_j)\}$$

for all $1 \le i \le \ell$ and $1 \le j \le s$. We compute $C_{\max}(i,j)$ row by row to obtain finally the total cost $C_{\max} = C_{\max}(\ell, s)$ of the maximum edit script. A similar procedure works for the minimum edit distance.

The key observation for us is to note that interdependency among the partial optimal weights $C_{\max}(i,j)$ induce an $\ell \times s$ grid-like directed acyclic graph, called further a *grid graph*. In such a graph vertices are points in the grid and edges go only from (i,j) point to

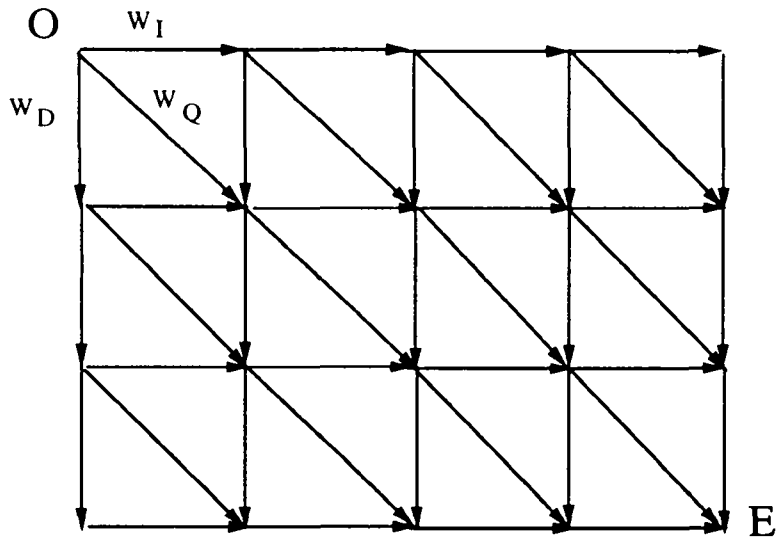

Figure 1: Example of a grid graph of size $\ell = 3$ and $s = 2$.

grid points $(i, j+1)$, $(i+1, j)$ and $(i+1, j+1)$. A horizontal edge from $(i, j-1)$ to (i, j) carries the weight $W_I(b_j)$; a vertical edge from $(i, j-1)$ to (i, j) has weight $W_D(a_i)$; and finally a diagonal edge from $(i-1, j-1)$ (i, j) is weighted according to $W_Q(a_i, b_j)$. Figure 1 shows an example of such an edit graph. The edit distance is the longest (shortest) path from the point $O = (0, 0)$ to $E = (\ell, s)$.

In this paper, we analyze the string edit problem in a probabilistic framework. We adopt the Bernoulli model for a random string, that is, *all symbols of a string are generated independently with probability p_i for symbol $i \in \Sigma$*. A standard probabilistic model assumes that both strings are generated according to the Bernoulli scheme (cf. [3], [6], [7], [8], [14], [22], [33], [34]). Such a model, however, leads to statistical dependency of weights in the associated grid graph. To avoid this problem, most of the time we shall work within the framework of another probabilistic model which postulates that only one string, say **b**, is random while the other is given (i.e., deterministic). This model has the advantage of having *independent* weights in the grid graph, and therefore, it is easier to analyze. We also show how to translate results of the latter model (called further the **independent model**) to the former one (called hereafter the **string model**). In passing we note that the independent model might be useful in some applications (e.g., when comparing a given string to all strings in a data base).

Clearly, in the independent model the distributions of weights $W_D(a_i)$, $W_I(b_j)$ and $W_Q(a_i, b_j)$ depend on the given string **a**. However, to avoid complicated notations we ignore this fact – whenever the independent model is discussed – and consider a grid graph

with weights W_I, W_D and W_Q. In other words, we concentrate on finding the longest path in a grid graph with independent weights W_I, W_D and W_Q, not necessary equally distributed. By selecting properly these distributions, we can model the string edit problem. For example, in the standard setting the deletion and insertion weights are identical, and usually constant, while the insertion weight takes two values, one (high) when matching between a letter of **a** and a letter of **b** occurs, and another value (low) in the case of a mismatch (e.g., in the *Longest Common Substring* problem, one sets $W_I = W_D = 0$, and $W_Q = 1$ when a matching occurs, and $W_Q = -\infty$ in the other case).

Our results for both models can be summarized as follows: Applying the *Subadditive Ergodic Theorem* we note that $C_{\max} \sim \alpha n$ almost surely (a.s.), where $n = \ell + s$ (cf. Theorem 2.1 and Corollary 2.2). Our main contribution lies in establishing bounds for the constant α (cf. Theorem 2.7) for the independent model, which by Corollary 2.2 can easily be extended to the string model. The upper bound is rather tight as verified by simulation experiments. More importantly, using the powerful and modern method of bounded differences (cf. [27]) we establish a sharp concentration of $C_{\max}$ around the mean value $EC_{\max}$ under a mild condition on the tail of the weight distributions (cf. Theorem 2.3). This proves the conjecture of Chang and Lampe [13] who observed empirically such a sharp concentration of $C_{\max}$ for a version of the string edit problem, namely the approximate string matching problem.

Our probabilistic results are proved in a unified manner by applying techniques of random walks (cf. [18], [20]), generating functions (cf. [19], [25], [26]), and bounded differences (cf. [27]). In fact, these techniques allow us to establish further results of a more general interest. In particular, we present a precise asymptotic estimate for the number of paths in the grid graph (cf. Theorem 2.4), which coincides with the number of sequence alignments (cf. [15], [16]). Finally, for the independent model we establish the limiting distribution of the total weight (cf. Theorem 2.5) and the tail distribution of the total weight (cf. Theorem 2.6) of a randomly selected path (edit script) in the grid graph.

The string edit problem and its special cases (e.g., the longest common subsequence problem and the approximate pattern matching) were studied quite extensively in the past, and are subject of further vigorous research due to their vital application in molecular biology. There are many algorithmic solutions to the problem, and we only mention here Apostolico and Guerra [1], Apostolico *et al.* [2], Chang and Lampe [13], Myeres [28], Ukkonen [32], and Waterman [34]. On the other hand, a probabilistic analysis of the problem was initiated by Chvatal and Sankoff [14] who analyzed the longest common subsequence problem. After an initial success in obtaining some probabilistic results for this problem, and its extensions by a rather straightforward applications of the subadditive ergodic theorem,

a deadlock was reached due to a strong interdependency between weights in the grid graph. To the best of our knowledge, there is no much literature on the probabilistic analysis of the string edit problem with a notable exception of a recent marvelous paper by Arratia and Waterman [7] (cf. [33]) who proved their own conjecture concerning phase transitions in a sequence matching.

There is, however, a substantial literature on probabilistic analysis of pattern matching. We mention here a series of papers by Arratia and Waterman (cf. [5], [6]) and with Gordon (cf. [3], [4]), as well as papers by Karlin and his co-authors (cf. [11], [21], [22]). Another approach for the probabilistic analysis of pattern matching with mismatches was recently reported by Atallah *et al.* in [8].

This paper is organized as follows. In the next section, we present our main results and discuss some of their consequences. Most of our proofs appear in Section 3.

2. MAIN RESULTS

In this section we present our main results concerning the typical behavior of the edit distance and the longest (shortest) path in a grid graph. We also report some findings on the probabilistic behavior of a randomly selected path in such a graph. We implicitly assume the independent model whenever the grid graph model is considered.

To recall, we consider a grid graph of size ℓ and s ($\ell \geq s$) as shown in Figure 1. All of our results, however, will be expressed in terms of $n = \ell + s$ and $d = \ell - s$. We assign to every edge in such a graph a real number representing its weight. (As mentioned in the Introduction, we ignore for the moment the fact that the weights depend on the string **a**.) A family of such directed acyclic weighted graphs will be denoted by $\vec{\mathcal{G}}(n,d)$ or shortly $\vec{\mathcal{G}}(n)$. We write $\vec{G}(n) \in \vec{\mathcal{G}}(n,d)$ for a member of such a family.

For the independent model we assume that *weights are independent* from an edge to an edge. Let $F_I(\cdot)$, $F_D(\cdot)$ and $F_Q(\cdot)$ denote distribution functions of W_I, W_D and W_Q respectively. We assume that the mean values m_I, m_D and m_Q, and the variances s_I^2, s_D^2 and s_Q^2, respectively, are finite. The distribution functions are not necessary identical.

The edit distance can be viewed as an optimization problem on the grid graph. Indeed, let $\mathcal{B}(n,d)$ or shortly $\mathcal{B}(n)$ be the set of all directed paths from the point O of the grid graph to the end point E. (It corresponds, as we know, to a script in the original string edit problem.) The cardinality of $\mathcal{B}(n)$, that is, the total number of paths between O and E, is denoted as $L(n,d)$. A particular path from O to E is denoted as $\mathcal{P}$, i.e., $\mathcal{P} \in \mathcal{B}(n,d)$. Note that the length $|\mathcal{P}|$ of a path $\mathcal{P}$ satisfies $\ell \leq |\mathcal{P}| \leq l + r = n$. Finally, let $N_I(\mathcal{P})$, $N_D(\mathcal{P})$ and $N_Q(\mathcal{P})$ denote the number of horizontal edges (say I-steps), vertical edges (say D-steps),

and diagonal edges (say Q-steps) in a path $\mathcal{P}$.

With the above notation in mind, the problem at hand can be posed as follows:

$$C_{\max} = \max_{\mathcal{P}\in\mathcal{B}(n)}\{W_n(\mathcal{P})\}\,, \tag{1}$$

$$C_{\min} = \min_{\mathcal{P}\in\mathcal{B}(n)}\{W_n(\mathcal{P})\} \tag{2}$$

where $W_n(\mathcal{P})$ denotes the total weight of the path $\mathcal{P}$ which becomes

$$W_n(\mathcal{P}) = \sum_{i=1}^{N_I(\mathcal{P})} W_I(i) + \sum_{i=1}^{N_D(\mathcal{P})} W_D(i) + \sum_{i=1}^{N_Q(\mathcal{P})} W_Q(i)\,. \tag{3}$$

We write W_n to denote the total weight of a randomly selected path. More precisely, we define

$$\Pr\{W_n < x\} = \frac{1}{L(n,d)}\sum_{\mathcal{P}\in\mathcal{B}} \Pr\{W_n(\mathcal{P}) < x\}\,. \tag{4}$$

Our results crucially depend on the order of magnitude of d with respect to n. We consider separately several cases, as defined below:

CASE (A): $d = O(\sqrt{n})$, and let $x = d\sqrt{\sqrt{2}/n} = \zeta d/\sqrt{n}$ where $\zeta = 2^{1/4}$.

CASE (B): $d = \Theta(n)$, and let $x = d/n$.

CASE (C): $d = n - O(n^{1-\epsilon})$, that is, for some constant x we have $d = n(1 - x/n^{\epsilon})$.

CASE (D): $d = O(1)$ (we shall reduce this case to Case (A)).

CASE (E): $s = O(1)$ (we shall reduce this case to case (C)).

Since the last two cases, as will turn out, can be reduced to the previous three ones, we shall concentrate below on the three top cases.

Now, we are in a position to present our results. To simplify further our presentation, we concentrate mainly on the longest path $C_{\max}$. We start with a simple general result concerning the typical behavior of $C_{\max}$. The more refined results containing a computable upper bound for $EC_{\max}$ (in the independent model) is given at the end of this section (cf. Theorem 2.7).

Theorem 2.1. *By the Subadditive Ergodic Theorem the following holds*

$$\lim_{n\to\infty}\frac{C_{\max}}{n} = \lim_{n\to\infty}\frac{EC_{\max}}{n} = \alpha \quad (a.s.)\,, \tag{5}$$

provided ℓ/s has a limit as $n \to \infty$. ■

The above result is true for both the independent and the string models. However, in the string model one has to consider the fact that weights depend on the given string **a**.

Let $P(\mathbf{a})$ be the probability of $\mathbf{a}$ occurrence in our standard Bernoulli model (e.g., for the binary alphabet $\Sigma = \{\alpha, \beta\}$ we have $P(\mathbf{a}) = p^{|\alpha|}(1-p)^{|\beta|}$ where p is the probability of α occurrence, and $|\alpha|$ ($|\beta|$) is the number of α's (β's) in the string $\mathbf{a}$). Since now the weights are functions of $P(\mathbf{a})$, hence the constant α in Theorem 2.1 (cf. (5)) depends on $\mathbf{a}$, too. Denote it by $\alpha_{\mathbf{a}}$. The following corollary follows directly from Theorem 2.1.

Corollary 2.2. *In the string model, the edit distance $C_{\max}$ satisfies (5) with α as below*

$$\alpha = \sum_{\mathbf{a} \in \mathcal{H}} \alpha_{\mathbf{a}} P(\mathbf{a}) \tag{6}$$

where $\mathcal{H}$ is the set of all possible strings $\mathbf{a}$ of length s over the alphabet Σ. ■

Finally, for both probabilistic models we can report the following finding concerning the concentration of the edit distance. It proves the conjecture of Chang and Lampe [13]. The proof of this result uses a powerful method of bounded differences (cf. [27]) (or in other words: Azuma's type inequality).

Theorem 2.3. (i) *If all weights are bounded random variables, say* $\max\{W_I, W_D, W_Q\} \leq 1$, *then for arbitrary $\varepsilon > 0$ and large n*

$$\Pr\{|C_{\max} - EC_{\max}| > \varepsilon EC_{\max}\} \leq 2\exp(-\varepsilon^2 \alpha n) \ . \tag{7}$$

(ii) *If the weights are unbounded but such that for large n, $W_{max} = \max\{W_I, W_D, W_Q\}$ satisfies the following*

$$n\Pr\{W_{max} \geq n^{1/2-\delta}\} \leq U(n) \tag{8}$$

for some $\delta > 0$ and a function $U(n) \to 0$ as $n \to \infty$, then

$$\Pr\{|C_{\max} - EC_{\max}| > \varepsilon EC_{\max}\} \leq 2\exp(-\beta n^{\delta}) + U(n) \tag{9}$$

for any $\varepsilon > 0$ and some $\beta > 0$.

Proof: Part (i) is a direct consequence of the following inequality of Azuma's type (cf. [27]): *Let X_i be i.i.d. random variables such that for some function $f(\cdot, \ldots, \cdot)$ the following is true*

$$|f(X_1, \ldots, X_i, \ldots, X_n) - f(X_1, \ldots, X_i', \ldots, X_n)| \leq c_i \ , \tag{10}$$

where $c_i < \infty$ are constants, and X_i' has the same distribution as X_i. Then,

$$\Pr\{|f(X_1, \ldots, X_i, \ldots, X_n) - Ef(X_1, \ldots, X_i, \ldots, X_n)| \geq t\} \leq 2\exp(-2t^2 / \sum_{i=1}^{n} c_i) \tag{11}$$

for some $t > 0$. The above technique is also called the method of bounded differences.

Now, for part (i) it suffices to set $X_i = b_i$ for $1 \le i \le \ell$, and $X_i = a_i$ for $\ell + 1 \le i \le n$, where a_i and b_i are the i symbols of the two strings **a** and **b**. Under our Bernoulli model, the X_i are i.i.d. and (10) holds, hence we can apply (11). Inequality (7) follows from the above and $t = \varepsilon EC_{\max} = O(n)$.

To prove part (ii), we note that for the string edit problem

$$|C_{\max}(X_1, \ldots, X_i, \ldots, X_n) - C_{\max}(X_1, \ldots, X_i', \ldots, X_n)| \le \max_{1 \le i \le n} \{X_i\} . \tag{12}$$

Then, for some constant c

$$\begin{aligned} \Pr\{|C_{\max} - EC_{\max}| \ge t\} &= \Pr\{|C_{\max} - EC_{\max}| \ge t \ , \ \max_{1 \le i \le n} \{X_i\} \le c\} \\ &+ \Pr\{|C_{\max} - EC_{\max}| \ge t \ , \ \max_{1 \le i \le n} \{X_i\} > c\} \\ &\le 2\exp(-2t^2/nc^2) + n\Pr\{X_i > c\} . \end{aligned}$$

Set now $t = \varepsilon EC_{\max} = O(n)$ and $c = O(n^{1/2-\delta})$, then

$$\Pr\{|C_{\max} - EC_{\max}| \ge \varepsilon EC_{\max}\} \le 2\exp(-\beta n^{\delta}) + n\Pr\{X_i > n^{1/2-\delta}\} ,$$

for some constant $\beta > 0$, and this implies (9) provided (8) holds. ∎

Hereafter, we assume only the independent model for which we have also obtained several new results regarding the probabilistic behavior of randomly selected path in a weighted grid graph. These results are of their own interests. For example, the total number of paths $L(n, d)$ in the grid graph (cf. Theorem 2.5) represents the number of ways the string **a** can be transformed into **b**, and this problem was already tackled by others (cf. [15], [16]) due to some applications in molecular biology. Furthermore, these results can be used to obtain refined probabilistic analysis of the string edit problem (cf. Theorem 2.7).

We start with the limiting distribution of the total weight of a randomly selected path.

Theorem 2.4. *The limiting distribution of the total weight is normal. More precisely,*

$$\frac{W_n - n\mu_W}{\sqrt{n}\sigma_W} \to \mathcal{N}(0, 1) \tag{13}$$

where $\mathcal{N}(0, 1)$ *is the standard normal distribution, and*

$$\mu_W = m_I\mu_I + m_D\mu_D + m_Q\mu_Q , \tag{14}$$

$$\sigma_W^2 = \mu_I s_I^2 + \mu_D s_D^2 + \mu_Q s_Q^2 + \tilde{\sigma}_Q^2(m_I + m_D - m_Q)^2 \tag{15}$$

where $\mu_I = EN_I(\mathcal{P})$, $\mu_D = EN_D(\mathcal{P})$, $\mu_Q = EN_Q(\mathcal{P})$ *and* $\tilde{\sigma}_Q^2 = varN_Q(\mathcal{P})$. *Explicit formulas for these quantities varies, and are given for each case (A)-(E) separately in the next section.* ■

The formulation of the next result concerning the number of paths in a grid graph $\vec{G}$ depends on a parameter u that takes different values for every case (A)-(E). Let us define $u = d/n$, and then x for every case (A)-(E) becomes:

CASE (A): Set $d = x\sqrt{n/\sqrt{2}}$. Then, $u = x/\sqrt{\sqrt{2}n} = x/(\zeta\sqrt{n})$.

CASE (B): Set $u = x$.

CASE (C): Set $d = n(1 - x/n^{\varepsilon})$. Then, $u = 1 - x/n^{\varepsilon}$.

CASE (D): Falls under (A).

CASE (E): As in case (C) with $x = 2s$ and $\varepsilon = 1$.

Then, we have the following result.

Theorem 2.5. *Let* $L(u) = L(n, d)$ *be the number of paths in a grid graph* $\vec{G} \in \vec{\mathcal{G}}(n)$, *where* u *is defined for each cases (A)-(E) as above. Then,*

$$L(u) = \frac{C\psi_2(\beta_2(u))^n}{\beta_2(u)^{n(1+u)/2}\sqrt{2\pi n V(u)}}(1 + O(1/n)) \tag{16}$$

where

$$\beta_2(u) = \frac{1 + 3u^2 + u\sqrt{8(u^2+1)}}{1 - u^2}, \tag{17}$$

$$\psi_2(u) = \psi_2[\beta_2(u)] = \frac{2u\beta_2(u)}{\beta_2(u) - 1 - u(1 + \beta_2(u))}, \tag{18}$$

and C *is a constant that is found in Section 3 (cf. (94)). In the above,* $V(u)$ *is the variance obtained from the generating function* $h(z)$ *defined as* $h(z) = \psi_2(z\beta_2(u))/\psi_2(\beta_2(u))$, *that is,* $V(u) = h''(1) - 0.25(1 - u^2)$ *where* $h''(z)$ *is the second derivative of* $h(z)$. ■

For most of our computations, we only need the asymptotics of $L(u)$ in the following rough form

$$\log L(u) = n\rho(u) - 0.5\log n + O(1), \tag{19}$$

where $\rho(u)$ differs for every case (A)-(E). In particular, for cases (A), (B) and (C) we have respectively

$$\rho(u) = -\log(\sqrt{2} - 1), \tag{20}$$

$$\rho(u) = \log\psi_2(\beta_2(u)) - \frac{1+u}{2}\log\beta_2(u), \tag{21}$$

$$\rho(x) = \frac{x\log n}{4\sqrt{n}} + \frac{x(1 + \log 4)}{2\sqrt{n}} - \frac{x\log x}{2\sqrt{n}} + O(\frac{1}{n}). \tag{22}$$

and x is defined above for the case (C). The details of the above derivations can be found in Section 3.

Finally, in order to obtain an upper bound for the cost $C_{\max}$, we need an estimate on the tail distribution of the total weight W_n along a random path. Formula (3) suggests to apply Cramer's large deviation result (cf. Feller [18]) with some modifications (due to the fact that the total weight W_n as in (3) is a sum of *random* number of weights). To avoid unnecessary complications, we consider in details only two cases, namely:

(a) all weights are *identically* distributed with mean $m = m_I = m_D = m_Q$ and the *cumultant function* $\Psi(s) = \log Ee^{s(W-m)}$ for the common weight $W - m$;

(b) insertion weight and deletion weight are constant, say all equal to -1 (e.g., $W_I = W_D = -1$), and the substitution weight $W_Q - m_Q$ has the cumultant function $\Psi_Q(s) = \log Ee^{s(W_Q - m_Q)}$. Such an assignment of weights is often encountered in the string edit problem.

In passing we note that the approach used in Section 3 to prove the next result, can handle also nonidentically distributed weights, however, algebra involved is quite tedious.

We prove the following result.

Theorem 2.6. (i) *In the case (a) of all identical weights, define s^* as the solution of*

$$a = \Psi'(s^*) \ , \tag{23}$$

for a given $a > 0$, and let

$$Z_0(a) = s^*\Psi'(s^*) - \Psi(s^*) \ , \tag{24}$$

$$E_1(a) = -(s^*m + \Psi(s^*)) \ , \tag{25}$$

$$E_2^2(a) = \frac{\tilde{\sigma}_Q^2 m^2 + 2\tilde{\sigma}_Q^2 ma + \tilde{\sigma}_Q^2(\Psi'(s^*))^2 + (1-\mu_Q)\Psi''(s^*)}{2(1-\mu_Q)\tilde{\sigma}_Q^2\Psi''(s^*)} \ , \tag{26}$$

where $\mu_Q = EN_Q$ and $\tilde{\sigma}_Q^2 = var\ N_Q$. Then,

$$\Pr\{W_n > (1-\mu_Q)(a+m)n\} \sim \frac{1}{2s^*E_2(a)\tilde{\sigma}_Q\sqrt{\pi(1-\mu_Q)n\Psi''(s^*)}} \exp\left(-n(1-\mu_Q)Z_0(a) + n\frac{E_1^2(a)}{4E_2^2(a)}\right) \tag{27}$$

(ii) *In case (b) of constant I-weights and D-weights, we define s^* as a solution of*

$$a = \Psi_Q'(s^*) \ , \tag{28}$$

and let

$$Z_0(a) = s^*\Psi'_Q(s^*) - \Psi_Q(s^*) , \tag{29}$$

$$E_1(a) = s^*(m_Q + 2) + 2s^*a^* - \Psi(s^*) , \tag{30}$$

$$E_2^2(a) = \frac{\tilde{\sigma}_Q^2(m_Q + 2)(m_Q + 2 + 2a^* + 4s^*\Psi''_Q(s^*)) + \tilde{\sigma}_Q a^*(a^* + 4s^*\Psi''_Q(s^*)) + \mu_Q\Psi''_Q(s^*)}{2\mu_Q\tilde{\sigma}_Q^2\Psi''_Q(s^*)} \tag{31}$$

Then,

$$\Pr\{W_n > \mu_Q(a + \beta/\mu_Q)n\} \sim \frac{1}{2s^*E_2(a)\tilde{\sigma}_Q\sqrt{\pi\mu_Q n\Psi''_Q(s^*)}} \exp\left(-n\mu_Q Z_0(a) + n\frac{E_1^2(a)}{4E_2^2(a)}\right) . \tag{32}$$

where $\beta = 2\mu_Q + m\mu_Q - 1$. ■

Having the above estimates on the tail of the total cost of a path in the grid graph $\vec{G} \in \vec{\mathcal{G}}(n)$, we can provide a more precise information about the constant α in our Theorem 2.1, that is, we compute an upper bound $\overline{\alpha}$ and a lower bound $\underline{\alpha}$ of α for the independent model. This bound can be used in Corollary 2.2 to obtain a bound for the constant in the string model. We prove below the following result, which is one of our main finding.

Theorem 2.7 *Assume the independent model.*
(i) *Consider first the identical weights case (cf. case (a) above). Let* a^* *be a solution of the following equation*

$$(1 - \mu_Q)Z_0(a^*) = \rho + \frac{E_1^2(a^*)}{4E_2^2(a^*)} , \tag{33}$$

where ρ *is defined in (20)-(22), and* Z_0, E_1 *and* E_2^2 *are defined in (24)-(26). Then, the upper bound* $\overline{\alpha}$ *of* α *becomes*

$$\overline{\alpha} = (1 - \mu_Q)(a^* + m) + O(\log n/n) . \tag{34}$$

In the case of constant I and D weights, let a^* *be a solution of the equation*

$$\mu_Q Z_0(a^*) = \rho + \frac{E_1^2(a^*)}{4E_2^2(a^*)} , \tag{35}$$

where Z_0, E_1 *and* E_2^2 *are as in (29-32). Then,*

$$\overline{\alpha} = \mu_Q(a^* + \beta/\mu_Q) + O(\log n/n) , \tag{36}$$

where β *is defined in Theorem 2.6(ii).*

(ii) *The lower bound* $\underline{\alpha}$ *of* α *can be obtained from a particular solution to our optimization problem (1). In particular, we have*

$$\underline{\alpha} = \max\{\mu_W, \ell m_D + s m_I, \alpha_{gr}\} \ , \tag{37}$$

where α_{gr} *is constructed from a greedy solution of the problem, that is,*

$$n\alpha_{gr} = (\ell + sp)m_{max} \tag{38}$$

where $p = \Pr\{W_Q > W_I \text{ and } W_Q > W_D\}$, *and* $m_{max} = E\max\{W_I, W_D, W_Q\}$.

Proof. We first prove part (i) provided Theorem 2.6 is *granted* (cf. Section 3 for the proof). Observe that by Boole's inequality we have for any real x

$$\Pr\{C_{\max} > x\} \le \sum_{\mathcal{P}\in\mathcal{B}} \Pr\{W_n(\mathcal{P}) > x\} = L(u)\Pr\{W_n > x\}$$

where the last equality follows from (4). Let now a_n be such that

$$L(u)\Pr\{W_n > a_n\} = 1 \ . \tag{39}$$

Then, in view of Theorem 2.6 one immediately proves that for $a_n \sim (1+\varepsilon)n\overline{\alpha}$ for any $\varepsilon > 0$, with $\overline{\alpha}$ as in (34) and (36), we have $\Pr\{C_{\max} > a_n\} = o(1)$, as needed for (34).

The lower bound can be established either by considering some particular paths $\mathcal{P}$ or applying a simple algorithm like a greedy one. The greedy algorithm selects in every step the most expensive edge, that is, the average cost per step is $m_{max} = E\max\{W_D, W_I, W_Q\}$. Let $p = \Pr\{W_Q > W_I \ , \ W_Q > W_D\}$. Then, assuming we have k D-steps, the number of Q-steps is binomially distributed with parameters p and $s-k$. Ignoring boundary conditions, we conclude that the average total cost for the greedy algorithm is

$$n\alpha_{gr} = m_{max}\sum_{k=1}^{s}(\ell + k)\binom{s}{k}p^{s-k}(1-p)^k = \ell + sp \ .$$

This formula should be modified accordingly if some boundary conditions must be taken into account. This can be accomplished in the same manner as discussed in Section 3. ∎

We compared our bounds for $C_{\max}$ with some simulation experiments. In the simulation we restricted our analysis to uniformly and exponentially distributed weights, and here we only report the latter results.

From Table 1 one concludes that the upper bound is quite tight. It is plausible that the normalized limiting distribution for $C_{\max}$ is double exponential (i.e., $e^{-e^{-x}}$), however, the normalizing constants are quit hard to find.

Table 1: Simulation results for exponentially distributed weights with means $m_I = m_D = m_Q = 1$ for case (B) with $d = 0.6n$.

ℓ	s	$\underline{\alpha}$	α_{sim}	$\overline{\alpha}$
200	50	1.588	1.909	2.45
400	100	1.588	1.808	2.45
600	150	1.5888	1.899	2.45
800	200	1.588	1.926	2.45
1000	250	1.588	1.922	2.45

The edit problem can be generalized, as it was recently done by Pevzer and Waterman [30] for the longest common subsequence problem. In terms of the grid graph, their generalization boils down to adding new edges in the grid graph that connect *no-neighboring* vertices. In such a situation our Theorem 2.1 may not hold, as easy to see. In fact, based on recent results of Newman [29] concerning the longest (unweighted) path in a general acyclic graph, we predict that a phase transition can occur, and $C_{\max}$ may switch from $\Theta(n)$ to $\Theta(\log n)$. This was already observed by Arratia and Waterman [7] for another string problem, namely, for the score in matching.

3. ANALYSIS THROUGH THE RANDOM WALK APPROACH

In this section, we only analyze the independent model. To recall, we consider a $\ell \times s$ grid graph with independent weights W_I, W_D and W_Q. To apply the random walk technique, we represent a path in the grid graph $\vec{G}$ as a spatial random walk. First of all, it is convenient to append our $\ell \times s$ graph to a full $\ell \times \ell$ grid graph, with all steps possible, as shown in Figure 2. It should be noted that in our new representation, a Q-step is twice as long as I-step and D-step.

At first, we analyze a path *without* weights in the grid graph shown in Figure 2. We call it an unweighted random walk (in short: R.W.) and denote as $Y(\cdot)$. Note that $Y(\cdot)$ is a spatial random walk that can move in the I-step direction of unit length, in the D-step direction of unit length, and in the Q-step direction of two units length. To model a path $\mathcal{P}$ in our original problem, we must assure that the random walk $Y(\cdot)$ coincides with the script path $\mathcal{P}$. This is achieved by putting a constraint on the random walk. More precisely, we require that the random walk $Y(\cdot)$ in Figure 2 ends at the point E of the grid graph after

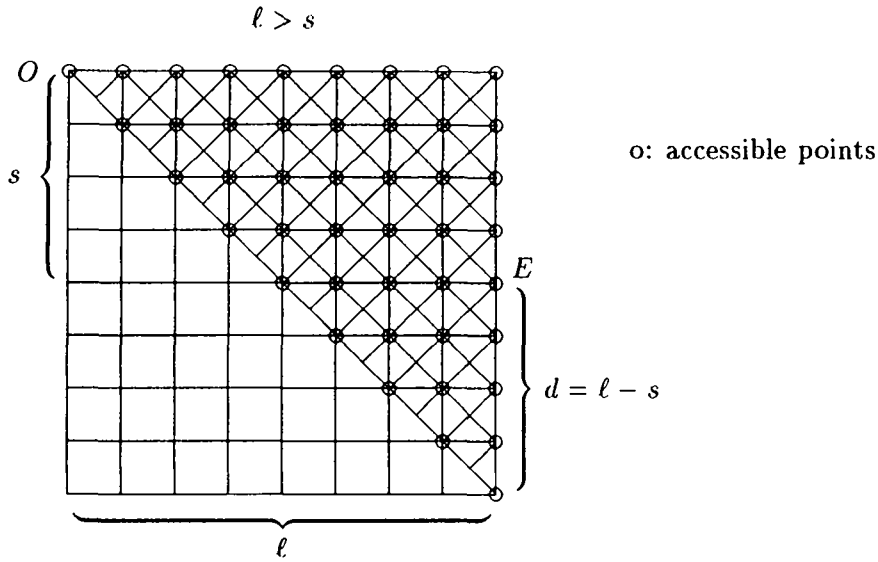

Figure 2: An extended $(\ell + 1) \times (\ell + 1)$ grid graph

n steps where $n = 2\ell - (\ell - s) = \ell + s$. In other words, the random walk in the $\ell \times \ell$ grid graph coincides with a script path if the following following constraint holds

$$Y(n) = d \tag{40}$$

where $d = \ell - s$.

To analyze $Y(\cdot)$, we first consider an *unconstraint* random walk $\hat{Y}(\cdot)$ such that the condition (40) does *not* hold, and that the probabilities of I-step, D-step and Q-step $\tau = \sqrt{2} - 1$, τ and τ^2 respectively, as shown in Figure 3. These probabilities are chosen in such a way that all paths with the same length receive the same probability (e.g., a two-step path ID has probability τ^2, the same as one-step path Q of length two).

The plan for this section is to analyze the large deviation behavior of the random walk $Y(n)$, and its weighted counterpart in order to assess the limiting distribution of the total weight. It turns out that the probabilistic behavior of $Y(n)$ depends on the relationship between d and n. We consider separately five cases (A)-(E) as discussed in the previous section. Fortunately, the last two cases can be reduced to the previous three ones, and they will not be discussed in details. Our ultimate goal is to show that the total weight of the weighted version of the random walk $Y(\cdot)$ is normally distributed as $n \to \infty$.

3.1 Case (A): $d = O(\sqrt{n})$

We first consider an *unconstraint* random walk $\hat{Y}(\cdot)$, that is, a random walk on the grid graph in Figure 2 without the constraint (40), and probabilities of move as in Figure 3. We

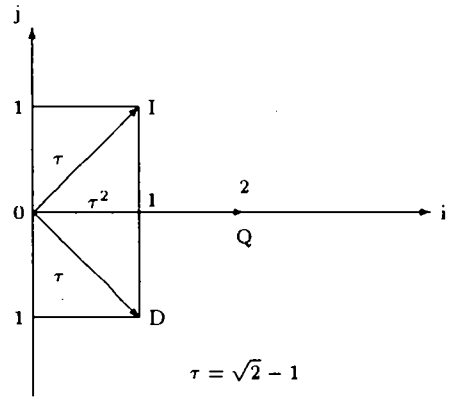

Figure 3: Probabilities of I-step, D-step and Q-step in the unconstraint random walk $\hat{Y}$

make the following scale changes $t = \frac{i}{n}$, $y = j\frac{\zeta}{\sqrt{n}}$ with $\zeta = \sqrt{\sqrt{2}}$ to establish the following theorem, where $\underset{n\to\infty}{\Rightarrow}$ represents the weak convergence of random functions in the space of all right continuous functions having left limit and endowed with the Skorohod metric (see Billingsley [9] Ch.III).

Theorem 3.1. *The unconstraint R.W. $\hat{Y}(\cdot)$ possesses the following limiting behavior*

$$\frac{\zeta \hat{Y}([nt])}{\sqrt{n}} \Rightarrow B(t), \qquad n \to \infty$$

where $B(\cdot)$ is a classical Brownian Motion (B.M.), and $\zeta = \sqrt{\sqrt{2}}$.

Proof. Let $p_i(j) = \Pr\{\hat{Y}(i) = j\}$. Then, from Figure 3 we infer the following

$$p_{i+1}(j) = \tau p_i(j-1) + \tau p_i(j+1) + \tau^2 p_{i-1}(j), \qquad i \geq 2 \tag{41}$$

Let $f(t,y)$ be the probability density in the new coordinate system, that is, we set $f(t,y)dy = p_{[nt]}([y\sqrt{n/\sqrt{2}}])$, where $[x]$ is an integer part of x. After subtracting $p_i(j)$ from (41), we obtain

$$\begin{aligned} \frac{1}{n}\partial_t f &= -\tau\frac{\zeta}{\sqrt{n}}\partial_y f + \frac{\tau\zeta^2}{2n}\frac{\partial^2 f}{\partial y^2} + \tau\frac{\zeta}{\sqrt{n}}\partial_y f \\ &+ \frac{\tau\zeta^2}{2n}\frac{\partial^2 f}{\partial y^2} - \frac{\tau^2}{n}\partial_t f + O\left(\frac{1}{n^{3/2}}\right) \end{aligned}$$

For $n \to \infty$, the above leads to the equation $\partial_t f = \frac{1}{2}\frac{\partial^2 f}{\partial y^2}$, which is exactly the heat equation characterizing the B.M. (cf. [18]). ■

Now, we take into account the constraint (40), that is, $Y(n) = d = O(\sqrt{n})$. Let

$$x = d\frac{\zeta}{\sqrt{n}} \tag{42}$$

where, to recall, $\zeta = 2^{1/4}$. To handle this constraint, we must recompute the probabilities of the I, D, and Q steps. We define these new one-step probabilities as follows

$$\begin{aligned} p_I &= \Pr\{\text{first move is } I |\ Y(n) = d\} \\ p_D &= \Pr\{\text{first move is } D |\ Y(n) = d\} \\ p_Q &= \Pr\{\text{first move is } Q |\ Y(n) = d\}\ . \end{aligned}$$

The new probabilities are computed in the lemma below.

Lemma 3.2a. *The new one-step probabilities become*

$$p_I = \tau\left(1 + \frac{\zeta}{\sqrt{n}}x\right) + O\left(\frac{1}{n}\right)\ , \tag{43}$$

$$p_D = \tau\left(1 - \frac{\zeta}{\sqrt{n}}x\right) + O\left(\frac{1}{n}\right)\ , \tag{44}$$

$$p_Q = \tau^2 + O\left(\frac{1}{n}\right)\ . \tag{45}$$

Proof. We know from Theorem 3.1 that $f(t,x) = \frac{e^{-\frac{x^2}{2t}}}{\sqrt{2\pi t}}$. This, and the above definitions of p_I, p_D and p_Q lead to the following

$$\begin{aligned} p_I &= \tau\frac{p_{i-1}(d-1)}{p_i(d)} \\ &= \frac{\tau}{p_i(d)}\left(p_i(\delta) - \frac{1}{n}\partial_t f - \frac{\zeta}{\sqrt{n}}\partial_x f + \frac{\zeta^2}{2n}\frac{\partial^2 f}{\partial x^2} + O\left(\frac{1}{n^{3/2}}\right)\right) \end{aligned}$$

But $\partial_x f = -\frac{x}{t}f$, hence p_I is now readily computed by setting $t = 1$ in the above. The two other probabilities are similarly derived. ∎

Remark 1. In fact, using similar arguments to the ones in the proof of Theorem 3.1, we can prove much stronger result. Namely, the constraint random walk $Y(\cdot)$ characterized by the probabilities p_I, p_D, p_Q has the limiting density given by $f(y,v) = \exp[-\frac{(y-x)^2}{2v}]/\sqrt{2\pi v}$, which is exactly the density of a B.M. with drift x. □

To estimate the limiting behavior of the total weight W_n as defined in (3), we need a precise evaluation of the random variables N_I, N_D and N_Q representing the number of I-steps, D-steps and Q-steps in a path $\mathcal{P}$. First of all, we compute the limiting distribution

of the sum $N_I + N_D + N_Q$. Using the renewal theory (cf. Feller [18], p. 321, 341, and Iglehart [20] Theorem 4.1) we can easily prove that

$$N_I + N_D + N_Q \sim \mathcal{N}\left(\frac{n}{\bar{d}}, n\frac{\bar{\sigma}^2}{\bar{d}^3}\right) + O(1), \qquad n \to \infty \tag{46}$$

where $\mathcal{N}(m, \sigma^2)$ is a classical Gaussian variable with mean m and variance (VAR) σ^2. In the above, $\bar{d}$ is the average *move step*, that is, from Lemma 3.2 we have

$$\bar{d} = p_I + p_D + 2p_Q = 1 + p_Q \ , \tag{47}$$

so that $\bar{d} = 2(2 - \sqrt{2}) + O(1/n)$, and

$$\bar{\sigma}^2 = p_Q(1 - p_Q) \tag{48}$$

so that $\bar{\sigma}^2 = \sqrt{2}(10 - 7\sqrt{2}) + O(1/n)$. Let

$$\alpha = \frac{1}{\bar{d}} = \frac{1}{1 + p_Q} = \frac{2 + \sqrt{2}}{4} + O\left(\frac{1}{n}\right) \tag{49}$$

and

$$\kappa = \frac{\bar{\sigma}^2}{\bar{d}^3} = \frac{\sqrt{2}}{16} + O\left(\frac{1}{n}\right) \tag{50}$$

Then, from (46), we obtain

$$N_I + N_D + N_Q \sim \mathcal{N}(n\alpha, n\kappa) + O(1)$$

From the expression (3) on the total weight W_n, it should be clear that we need the joint distribution of N_I, N_D and N_Q. For this, we must consider two constraints on $N_\cdot$:[1] one on the total number of steps, and the other related to $Y(n) = d$. More precisely, together with (42) we have the following constraint on the number of steps

$$N_I + N_D + 2N_Q = n \tag{51}$$

$$N_I - N_D = d = x\frac{\sqrt{n}}{\zeta} \tag{52}$$

The above constraints must be taken into account to assess the joint limiting distribution of the number of steps. To accomplish this, we shall use a technique introduced in Louchard, Schott and Randrianarimanana [26].

We first consider only the constraint (51). This will allow us to compute the asymptotic joint distribution of N_I, N_D, N_Q, as stated in the next theorem.

[1] To simplify our notation, we often write $X_\cdot$ to denote any of X_I, X_D or X_Q.

Theorem 3.3a. *The number of I, D and Q steps, N_I, N_D, N_Q are asymptotically Gaussian, with mean $n\mu_I$, $n\mu_D$, $n\mu_Q$ respectively, where*

$$\mu_I = \bar{p}_I(2\alpha - 1) = p_I/(1 + p_Q) = \frac{\sqrt{2}}{4}\left(1 + \frac{\zeta}{\sqrt{n}}x + O(\frac{1}{n})\right) , \tag{53}$$

$$\mu_D = \bar{p}_D(2\alpha - 1) = p_D/(1 + p_Q) = \frac{\sqrt{2}}{4}\left(1 - \frac{\zeta}{\sqrt{n}}x + O(\frac{1}{n})\right) , \tag{54}$$

$$\mu_Q = (1 - \alpha) = p_Q/(1 + p_Q) = \frac{2 - \sqrt{2}}{4} + O(\frac{1}{n}) \tag{55}$$

where $\bar{p}_I = p_I/(p_I + p_D)$, $\bar{p}_D = p_D/(p_I + p_D)$. Moreover, α is given by (49). The asymptotic covariance matrix is given by

$$n \cdot \begin{matrix} I \\ D \\ Q \end{matrix} \begin{pmatrix} \sigma_I^2 & C_{ID} & -2\bar{p}_I\kappa \\ C_{ID} & \sigma_D^2 & -2\bar{p}_D\kappa \\ C_{IQ} & -2\bar{p}_D\kappa & \kappa \end{pmatrix} \tag{56}$$

with

$$\sigma_I^2 = (2\alpha - 1)\bar{p}_I(1 - \bar{p}_I) + 4\kappa\bar{p}_I^2 = \frac{3\sqrt{2}}{16} + \frac{2^{3/4}}{8}x + O(\frac{1}{n})$$

$$\sigma_D^2 = (2\alpha - 1)\bar{p}_D(1 - \bar{p}_D) + 4\kappa\bar{p}_D^2 = \frac{3\sqrt{2}}{16} - \frac{2^{3/4}}{8}x + O(\frac{1}{n})$$

$$C_{ID} = 2\kappa - (2\alpha - 1)\bar{p}_I\bar{p}_D - 2\kappa(\bar{p}_I^2 + \bar{p}_D^2) = \frac{\sqrt{2}}{16} + O(\frac{1}{n})$$

where κ is given in (50).

Proof. From (46) and (51), and setting $N_T = N_I + N_D$, we see that

$$\frac{n}{2} + \frac{N_T}{2} = n - N_Q \sim \mathcal{N}(n\alpha, n\kappa) + O(1) \tag{57}$$

$$\mu_T = E(N_T)/n = n(2\alpha - 1) + O(1) \tag{58}$$

$$\sigma_T^2 = \mathrm{VAR}(N_T)/n \sim 4\kappa$$

But *given* N_T, the number of I-steps N_I is a binomial random variable with parameter $\bar{p}_I$, and mean $N_T\bar{p}_I$ and the variance $N_T\bar{p}_I\bar{q}_I$ (where $\bar{q}_I = 1 - \bar{p}_I$). By (58) we have

$$E(N_I) = n\mu_T\bar{p}_I = n\bar{p}_I(2\alpha - 1)$$

We also obtain

$$E(N_I^2|N_T) = N_T\bar{p}_I\bar{q}_I + N_T^2\bar{p}_I^2 ,$$

and

$$E(N_I^2) = n\mu_T\bar{p}_I\bar{q}_I + \bar{p}_I^2(n\sigma_T^2 + n^2\mu_T^2) \ .$$

This finally leads to

$$\sigma_I^2 = n\left(\mu_T\bar{p}_I\bar{q}_I + \bar{p}_I^2\sigma_T^2\right) \sim n((2\alpha - 1)\bar{p}_I\bar{q}_I + \bar{p}_I^2 4\kappa)$$

The number of D-steps is analyzed in a similar fashion. To compute the covariance C_{ID} between N_I and N_D, note that

$$n\sigma_T^2 = \mathrm{VAR}(N_T) = \mathrm{VAR}(N_I + N_D) = n(\sigma_I^2 + \sigma_D^2 + 2C_{ID})$$

or $4\kappa \sim 2(2\alpha - 1)\bar{p}_I\bar{q}_I + 4\kappa(\bar{p}_I^2 + \bar{p}_D^2) + 2C_{ID}$. Finally,

$$\begin{aligned} COV(N_I N_T) &= E(N_I N_T) - E(N_I)E(N_T) \\ &= \bar{p}_I E(N_T^2) - E(N_I)E(N_T) = n\bar{p}_I\sigma_T^2 \sim \bar{p}_I 4\kappa n \end{aligned}$$

and with (57), we obtain $COV(N_I N_Q) = -\frac{1}{2}COV(N_I N_T) \sim -\bar{p}_I 2\kappa n$.

To complete the proof, it suffices to check the asymptotic Gaussian property of N_I, N_D, N_Q. For N_Q, this follows from (57). For N_I, which is binomially distributed with parameter $\bar{p}_I$, we obtain, conditioning on N_T

$$\begin{aligned} E[e^{i\theta N_I/\sqrt{n}}] &= E\left\{[\bar{p}_I e^{i\theta/\sqrt{n}} + \bar{q}_I]^{N_T}\right\} \\ &= E\left\{[1 + \bar{p}_I i\frac{\theta}{\sqrt{n}} - \bar{p}_I\frac{\theta^2}{2n} + O(\frac{1}{n^{3/2}})]^{N_T}\right\} \\ &= E\left\{\exp N_T[i\bar{p}_I\frac{\theta}{\sqrt{n}} - \frac{\theta^2}{2n}\bar{p}_I\bar{q}_I + O(\frac{1}{n^{3/2}})]\right\} \end{aligned} \tag{59}$$

But, by (57),

$$E[e^{i\rho N_T}] = e^{in\mu_T\rho - \frac{1}{2}n\sigma_T^2\rho^2 + O(\rho^3\frac{n^{3/2}}{\sqrt{n}})}$$

hence by (59) we obtain

$$E[e^{i\theta N_I/\sqrt{n}}] = \exp\left\{in\mu_T\frac{\theta}{\sqrt{n}}\bar{p}_I - \frac{\theta^2}{2}(\bar{p}_I\bar{q}_I\mu_T + \sigma_T^2\bar{p}_I^2) + O(\frac{1}{\sqrt{n}})\right\}$$

which proves the asymptotic Gaussian property of N_I. ■

To complete our study of the number of steps in the grid graph, we only need to consider the constraint (52). Set $\eta_\cdot = (N_\cdot - n\mu_\cdot)/\sqrt{n}$. Note that by Lemma 3.2a and Theorem 3.3a, $E(N_I - N_D) = \frac{\sqrt{n}}{\zeta}x + O(1)$ as it should be, so (52) implies that

$$\eta_I = \eta_D \ , \tag{60}$$

and , by (51),

$$\eta_I = -\eta_Q \ . \tag{61}$$

To derive the *constrained* density of η_Q, we first write the joint asymptotic density of (η_Q, η_I), which by Theorem 3.3a becomes

$$\frac{\exp\left\{-\frac{1}{2(1-R^2)}\left[\frac{\eta_I^2}{\sigma_I^2} - 2R\frac{\eta_I\eta_Q}{\sigma_I\sigma_Q} + \frac{\eta_Q^2}{\sigma_Q^2}\right]\right\}}{2\pi\sigma_I\sigma_Q\sqrt{1-R^2}} \tag{62}$$

with $R = \frac{C_{IQ}}{\sigma_I\sigma_Q}$.

Setting $\eta_I = -\eta_Q$, we finally obtain the asymptotic density of η_Q, as stated below.

Lemma 3.4a *Under constraint (52), we have*

$$\eta_Q \sim \mathcal{N}(0, \tilde{\sigma}_Q^2) + O(\frac{1}{\sqrt{n}}) \tag{63}$$

with

$$\begin{aligned} \tilde{\sigma}_Q^2 &= (1-R^2)\left(\frac{1}{\sigma_I^2} + 2\frac{C_{IQ}}{\sigma_I^2\sigma_Q^2} + \frac{1}{\sigma_Q^2}\right)^{-1} \\ &= \sqrt{2}/16 + O(1/n) \ , \end{aligned} \tag{64}$$

where all the quantities in the above were defined before. ■

We are now ready to analyze the total cost along a path $\mathcal{P}$ from O to E. For the reader convenience, we repeat below our formula (3) for the total weight W_n

$$W_n(\mathcal{P}) = \sum_{i=1}^{N_I(\mathcal{P})} W_I(i) + \sum_{i=1}^{N_D(\mathcal{P})} W_D(i) + \sum_{i=1}^{N_Q(\mathcal{P})} W_Q(i) \ . \tag{65}$$

The next theorem provides the limiting distribution for the total weight W_n, and it proves our presented Theorem 2.4 for the case $d = O(\sqrt{n})$. To simplify some of our notation, we often write $\sum_A \Psi_A = \Psi_I + \Psi_D + \Psi_Q$ for any random variable Ψ..

Theorem 3.5a *The total weight W_n is asymptotically Gaussian, with mean*

$$EW_n = n\mu_W = n\sum_A m_A\mu_A \tag{66}$$

and

$$\sigma_W^2 = \frac{\mathrm{VAR}(W_n)}{n} = \sum_A \mu_A s_A^2 + \tilde{\sigma}_Q^2(m_I + m_D - m_Q)^2 \ .$$

Proof. The mean EW_n is clearly given by

$$EW_n = n \sum_A m_A \mu_A$$

Now, by (60) and (61) all random variables can be written in term of η_Q. We have

$$\begin{aligned}
\mathrm{VAR}(W_n) &= EW_n^2 - (EW_n)^2 \\
&= E\left(\sum_A N_A E(W_A^2) + \sum_A N_A(N_A-1)m_A^2 + 2\sum\sum_{A\neq B} N_A N_B m_A m_B \right) - (EW_n)^2 \\
&= n\left\{ \sum_A \mu_A E(W_A^2) + \sum_A ((\tilde{\sigma}_Q^2 + n\mu_A^2) - \mu_A)m_A^2 \right. \\
&+ 2(\tilde{\sigma}_Q^2 + n\mu_I\mu_D)m_I m_D + 2(-\tilde{\sigma}_Q^2 + n\mu_I\mu_Q)m_I m_Q \\
&+ \left. 2(-\tilde{\sigma}_Q^2 + n\mu_D\mu_Q)m_D m_Q \right\} - (EW_n)^2
\end{aligned}$$

After some simplification, we obtain

$$\mathrm{VAR}(W_n) = n\left(\sum_A \mu_A s_A^2 + \tilde{\sigma}_Q^2 (m_I + m_D - m_Q)^2 \right)$$

For the limiting density, we first observe that $N_\cdot$ are all $O(n)$, which allows us to use the *Central Limit Theorem* for each component of W_n. This leads to

$$\begin{aligned}
E[e^{i\theta W_n/\sqrt{n}}] &= E\left\{ \exp\left[\sum_A \left(-\frac{1}{2}s_A^2 \frac{\theta^2}{n} + \frac{i m_A \theta}{\sqrt{n}} + O(\frac{1}{n^{3/2}}) \right) N_A \right] \right\} \\
&= E\left\{ \exp\left[(-\frac{1}{2}s_I^2\frac{\theta^2}{n} + i\frac{m_I\theta}{\sqrt{n}})(n\mu_I - \sqrt{n}\eta_Q) + (-\frac{1}{2}s_D^2\frac{\theta^2}{n} + i\frac{m_D\theta}{\sqrt{n}})(n\mu_D - \sqrt{n}\,\eta_Q) \right.\right. \\
&+ \left.\left. (-\frac{1}{2}s_Q^2\frac{\theta^2}{n} + i\frac{m_Q\theta}{\sqrt{n}})(n\mu_Q + \sqrt{n}\,\eta_Q) + O(\frac{1}{\sqrt{n}}) \right] \right\} \\
&= \exp[-\frac{1}{2}\theta^2(\sum_A \mu_A s_A^2) + i\sqrt{n}\,\theta(\sum_A m_A\mu_A)] \cdot \\
&\quad \cdot E\left\{ \exp\left[\eta_Q i\theta(-m_I - m_D + m_Q) + O(\frac{1}{\sqrt{n}}) \right] \right\} \\
&= \exp\left[-\frac{1}{2}\theta^2(\sum_A \mu_A s_A^2) + i\sqrt{n}\,\theta(\sum_A m_A\mu_A) - \frac{1}{2}\tilde{\sigma}_Q^2\theta^2(m_I + m_D - m_Q)^2 + O(\frac{1}{\sqrt{n}}) \right]
\end{aligned}$$

which completes the proof. ■

We delay the discussion of the number of paths $L(n,d)$ (cf. Theorem 2.5) till the analysis of the case (B). It will turn out that the asymptotics of $L(n,d)$ for (A) can be deduced from the asymptotics of $L(n,d)$ obtained in case (B).

Finally, we prove our last result concerning the tail of the total weight distribution (cf. Theorem 2.6). As discussed in Section 2, we only consider two cases, namely: (a) identically distributed weights, that is, $W_I =^d W_D =^d W_Q = W$ where $=^d$ means equal *in distribution*; and (b) constant D-weight and I-weight, i.e., $W_D = W_I = -1$. The proof of Theorem 2.6 relies heavily on applying the large deviation results (cf. Bucklew [10] and Feller [18]), and we concentrate on proving the identical distribution case providing only sketchy explanations for the other case.

Let us first establish notation needed to express a large deviation result. Define $S_n = \sum_{i=1}^{n} W(i)$ where $W(i)$ is an independent copy of W. Let $\Psi(z) = \log Ee^{z(W-m)}$ be the cumulant function of $W - m$ where $m = EW$, and let s be the unique solution, if exists, of the following equation

$$a = \Psi'(s)$$

for any $a > 0$. Finally, let $Z(a) = -(\Psi(s) - s\Psi'(s))$. Then (cf. Feller [18])

$$\Pr\{S_n \geq n(a+m)\} \sim \frac{1}{s\sqrt{2\pi n\Psi''(s)}} \exp(-nZ(a)) \; . \tag{67}$$

In our case, the total weight $W_n(N_Q)$ of a random path in a grid graph with exactly N_Q diagonal edges becomes $W_n(N_Q) = \sum_{i=1}^{N_I} W_I(i) + \sum_{i=1}^{N_D} W_D(i) + \sum_{i=1}^{N_Q} W_Q(i) = \sum_{i=1}^{n-N_Q} W(i)$ (cf. (51)). Note that N_Q is a random variable, hence the unconditional total weight W_n can be computed from an estimate of the conditional total weight $W_n(N_Q)$ and the limiting distribution of N_Q (cf. Lemma 3.4a). But, $N_Q = n\mu_Q + \eta_Q\sqrt{n}$ and by Lemma 3.4a η_Q is asymptotically normal with mean 0 and variance $\tilde{\sigma}_Q^2$. We must now translate (67) into our new situation. Let $\tilde{n} = \gamma n$ where $\gamma = 1 - \mu_Q$. Define $\tilde{a}$ such that

$$\begin{aligned} \tilde{n}a + m\sqrt{n}\eta_Q &= (\tilde{n} - \sqrt{n}\eta_Q)\tilde{a}, \qquad \text{i.e.} \\ \tilde{a} &= a + \frac{(m+a)}{\gamma}\frac{\eta_Q}{\sqrt{n}} + \frac{m+a}{\gamma^2}\frac{\eta_Q^2}{n} + O\left(\frac{\eta_Q^3}{n^{3/2}}\right) \end{aligned}$$

Let also s^* and s be solutions of the following equations $a = \Psi'(s^*)$ and $\tilde{a} = \Psi'(s)$. Using Taylor's expansion of $\Psi(s)$ and $\Psi'(s)$ around s^*, we obtain

$$s = s^* + \frac{a^* + m}{\gamma\psi''(s^*)}\frac{\eta_Q}{\sqrt{n}} - \frac{1}{2}\frac{(a^*+m)[-2(\psi''(s^*))^2 + (a+m)\psi'''(s^*)]}{\gamma^2(\psi''(s^*))^3}\frac{\eta_Q^2}{n} + O(\frac{1}{n^{3/2}}) \; . \tag{68}$$

With the notation as above, we reduce the problem to the following one

$$\Pr\{W_n \geq \gamma(a+m)n\} = \int_{-\infty}^{\infty} \Pr\{\sum_{i=1}^{\tilde{n}-\sqrt{n}\eta} (W(i) - m) \geq (\tilde{n} - \sqrt{n}\eta)\tilde{a} | \eta_Q = \eta\} dF_{\eta_Q}(\eta)$$

where $F_{\eta_Q}(\eta) = \Phi(\eta)(1 + O(1/\sqrt{n}))$ (cf. Lemma 3.4a) with $\Phi(\cdot)$ denoting the distribution function of the normal distribution with mean zero and variance $\tilde{\sigma}_Q^2$. The probability under the above integral can be estimated as in (67). Using, in addition, the well known formula

$$\int_{-\infty}^{\infty} \exp(-p^2x^2 \pm qx)dx = \frac{\sqrt{\pi}}{p} \exp\left(\frac{q^2}{4p^2}\right) ,$$

after tedious algebra, we obtain our result (27) from Theorem 2.6.

In a similar manner we deal with the second case (b). However, this time the starting equation is $W_n(N_Q) = \sum_{i=1}^{N_Q} W_Q(i) - (n - 2N_Q)$. The details are left to the reader.

The above proof works line by line for all other cases (B)-(E), however, one must use appropriate value for $\tilde{\sigma}_Q^2$ (cf. (64) in Lemma 3.4a with new probabilities p_I, p_D and p_Q for other cases (B)-(E)).

3.2 Case (B): $d = O(n)$

The main purpose of this section is to derive the limiting distribution of the total weight for a given path $\mathcal{P}$ in a grid graph $\vec{G} \in \vec{\mathcal{G}}$. As in the previous subsection, we proceed in three steps: at first, we consider an unweighted unconstraint random walk, then we derive probabilities p_I, p_D and p_Q for the constraint unweighted random walk $Y(\cdot)$, and finally we deal with the total weight W_n.

Consider the unweighted random walk $Y(\cdot)$ in the grid graph as in Figure 2 such that $Y(n) = d = nx$ for some $x < 1$. Naturally, in this domain of d and n we can*not* use the normal approximation, which works only up to $O(\sqrt{n})$. We have to appeal to the large deviation arguments to obtain the probability distribution of the random walk $Y(\cdot)$. We proceed along the lines of arguments suggested by Louchard [25].

We consider the constraint random walk $Y(n) = nx$, however, for further analysis we must generalize our constraint to the following one

$$Y(m) = mu \ . \tag{69}$$

One can imagine that the random walk $Y(\cdot)$ at step m has to be at position mu, where m and u are functions of n and x (e.g., we shall assume that $mu = nx$).

As in the case (A), the analysis of the number of steps N_I, N_D and N_Q is crucial for the total weight. Note that, under our constraint (69), we have

$$\begin{aligned} N_I + N_D + 2N_Q &= m \\ N_I - N_D &= mu \end{aligned}$$

The above can be translated to the following constraint: $N_I + N_Q = \frac{m}{2}(1+u)$. Bearing this in mind, we transformed the random walk $Y(\cdot)$ into another random walk $\tilde{Y}(\cdot)$ that is defined in Figure 4 below (i.e., its one-step moves are shown in Fig. 4). Our interest lies

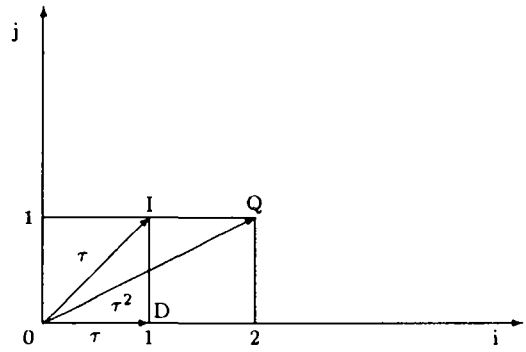

Figure 4: Definition of the new random walk $\tilde{Y}(\cdot)$.

in estimating $\Pr\{Y(m) \in mdu\}$ or in terms of the new random walk $\tilde{Y}(\cdot)$ we evaluate the following

$$\Pr\{Y(m) \in mdu\} \equiv \Pr\left\{\frac{\tilde{Y}(m) - m/2}{m} \in \frac{du}{2}\right\} \tag{70}$$

To analyze $\tilde{Y}(\cdot)$, we compute the probability $p_i(j) = \Pr\{\tilde{Y}(i) = j\}$. It is easy to see that this probability satisfies the following recurrence

$$p_{i+1}(j) = \tau p_i(j) + \tau p_i(j-1) + \tau^2 p_{i-1}(j-1)\ , \qquad i \geq 1\ . \tag{71}$$

We solve this recurrence by the mean of generating function approach. Let the generating function (G.F.) of $p_i(j)$ with respect to j be defined as below

$$g_i(z) = \sum_{j=0}^{\infty} z^j p_i(j)\ .$$

After multiplying recurrence (71) by z^j and summing up, we immediately arrive at

$$\begin{aligned} g_0(z) &= 1, \qquad g_1(z) = \tau(1+z) \\ g_{i+1}(z) &= \tau g_i(z) + \tau z g_i(z) + \tau^2 z g_{i-1}(z), \qquad i \geq 1 \end{aligned} \tag{72}$$

Let now the bivariate generating function $\varphi(\theta, z)$ of $g_i(z)$ be defined as

$$\varphi(\theta, z) = \sum_{i=0}^{\infty} \theta^i g_i(z)\ .$$

Our previous estimate (72) for $g_i(z)$ leads to

$$\varphi(\theta,z)-1-\theta g_1(z)=\tau\theta(\varphi(\theta,z)-1)+\tau\theta z(\varphi(\theta,z)-1)+\tau^2\theta^2 z\varphi(\theta,z)\ ,$$

which can be solved to obtain

$$\varphi(\theta,z)=\frac{1}{1-(1+z)\theta\tau-z\theta^2\tau^2} \tag{73}$$

The roots of the denominator of the above become

$$\theta_{1,2}(z)=-\frac{(1+z)\tau\pm\tau\sqrt{w_1(z)}}{2z\tau^2}$$

where $w_1(z)=1+6z+z^2$.

To simplify further our presentation, we will denote by f^* the value of any function $f(z)$ at $z=1$. In particular, we have

$$\theta_1^*=-(3+2\sqrt{2}),\qquad \theta_2^*=1 \tag{74}$$

Furthermore, our basic equation (73) can be transformed to

$$\varphi(\theta,z)=\left(\frac{\alpha_1(z)}{\theta-\theta_1(z)}+\frac{\alpha_2(z)}{\theta-\theta_2(z)}\right)\ . \tag{75}$$

where

$$\alpha_2(z)=-\frac{1}{\tau\sqrt{w_1(z)}},\qquad \alpha_1(z)=-\alpha_2(z)\ .$$

To extract the generating function $g_i(z)$ from (75), we expand $\varphi(\theta,z)$ in the powers of θ to obtain

$$g_m(z)=-\frac{\alpha_1(z)}{\theta_1(z)}\left(\frac{1}{\theta_1(z)}\right)^m-\frac{\alpha_2(z)}{\theta_2(z)}\left(\frac{1}{\theta_2(z)}\right)^m \tag{76}$$

Since we are interested in large values of m, we deduce from (76)

$$g_m(z)\sim\frac{1}{\theta_2(z)})^m=\psi_1^m(z),\qquad m\to\infty \tag{77}$$

with $\psi_1(z)=1/\theta_2(z)$.

Our aim now is to assess asymptotically the probability $p_m(k)=\Pr\{\widetilde{Y}(m)=k\}$. Clearly, it can be estimated as $p_m(k)\sim[\psi_1^m(z)]_k$ where $[f(z)]_k$ is the coefficient of z^k in the power expansion of $f(z)$. Hence, we have to deal with evaluating the kth coefficient of of $\psi_1^m(z)$, where $k=m(1+u)/2$. To obtain such asymptotics we shall use the classical "shift of the mean" technique (cf. Feller [18] p.548 and Greene and Knuth [19] p.79). For the reader convenience, we discuss briefly this technique below. We follow the approach of Greene and Knuth [19].

Let $g(z)$ be the generating function of a random variable with mean equal to μ and the variance equal to σ^2. Then, $g^n(z)$ represents the generating function of the sum of n such random variables. We estimate the coefficient of $z^{\mu n+r}$ in $g^n(z)$ for such r that $\mu n + r$ is an integer. Call such a coefficient $A_{n,r}$. By the Cauchy formula we have

$$A_{n,r} = \frac{1}{2\pi i} \oint \frac{g^n(z)dz}{z^{\mu n+r+1}} ,$$

where the integral is taken over a circle that encloses the origin. Applying the *saddle point method* Greene and Knuth derive the following formula

$$A_{n,r} = \frac{1}{\sigma\sqrt{2\pi n}} \exp\left(\frac{-r^2}{2\sigma^2 n}\right) + O(n^{3\varepsilon-1}) \tag{78}$$

where ε is arbitrary small positive number. The reader should notice that this asymptotics is valid *only* for $r = O(\sqrt{n})$.

In our case, we need the kth coefficient of $\psi_1^m(z)$, where $k = m(1+u)/2$. Therefore, we *cannot* directly apply (78) since we are not in the range $O(\sqrt{m})$. A solution to this dilema is proposed in [19] by a simple and elegant application of the "shift of the mean" technique, which we discuss below.

Let us return to the notation of Greene and Knuth [19], and assume one needs the k coefficient of $g^n(z)$. The shift of the mean technique computes the k coefficient as follows

$$[g^n(z)]_k = \frac{g(\beta)^n}{\beta^k} \left[\left(\frac{g(\beta z)}{g(\beta)}\right)^n\right]_k , \tag{79}$$

where the parameter β allows to shift the mean of the distribution to a value close to k/n, and hence allows to apply the asymptotics (78). The choice of β is specified by the following equation

$$\frac{\beta g'(\beta)}{g(\beta)} = \frac{k}{n} . \tag{80}$$

Now, we are ready to derive our asymptotics. Since we seek the $k = m(1+u)/2$ coefficient of $\psi_1^m(z)$, we first apply (80) to shift the mean. Define $\beta_1(u)$ as

$$\frac{\beta_1 \psi_1'(\beta_1)}{\psi_1(\beta_1)} = \frac{1+u}{2} \tag{81}$$

The probability given by (70) is now related to the new generating function $\psi_1(\beta_1 z)/\psi_1(\beta_1)$. Its mean is given by (81) and its variance (the exact value of which is of no interest for us) will be denoted by $V(u)$. Finally, applying (78), we obtain our main result.

Theorem 3.1b *We have proved*

$$\begin{aligned}\Psi_1(m,u) &= \Pr\{Y(m) \in mdu\} = \Pr\{\tilde{Y}(m) - \frac{m}{2} \in \frac{mdu}{2}\} \\ &\sim \frac{(\psi_1(\beta_1))^m}{\beta_1^{m(1+u)/2}\sqrt{2\pi m V(u)}} \frac{mdu}{2}(1+O(1/n))\end{aligned}$$

for all m and $u < 1$. ■

Let us now compute $\beta_1(u)$ and $\psi_1(\beta_1(u))$. After some algebra, using (80) and (81) we have

$$\frac{(\beta_1(u)-1)\psi_1(u)}{2\tau\beta_1(u) + (1+\beta_1(u))\psi_1(u)} = u \tag{82}$$

$$\beta_1(u) = \frac{1+3u^2+u\sqrt{8(u^2+1)}}{1-u^2} \tag{83}$$

$$\psi_1(u) = \psi_1[\beta_1(u)] = \frac{2u\tau\beta_1(u)}{\beta_1(u)-1-u(1+\beta_1(u))} \tag{84}$$

Theorem 3.1b allows to analyze the constraint random walk $Y(n) = d$. In particular, as for case (A), we can compute the probabilities p_I, p_D, and p_Q of one-step moves. Setting in Theorem 3.1b, $m = nt, u = \frac{x}{t}$ so that $mu = nx$, we obtain

$$\begin{aligned}\Psi_2(x,t)dx &= \Pr\{Y(nt) \in ndx\} \\ &\sim \exp\left\{nt[\log\psi_1(\frac{x}{t}) - \frac{1}{2}\log\beta_1(\frac{x}{t})] - \frac{nx}{2}\log\beta_1(\frac{x}{t})\right\} \frac{\sqrt{n}\,\lambda[\beta_1(\frac{x}{t})]}{2\sqrt{2\pi t V(\frac{x}{t})}}\,.\,dx\end{aligned}$$

This implies, for example, that

$$p_I \sim \frac{1}{3}\left[\frac{\Psi_2(x-\frac{1}{n}, t-\frac{1}{n})}{\Psi_2(x,t)}\right]_{t=1}$$

and in a similar fashion for D and Q. After some algebra, we finally derived the following lemma.

Lemma 3.2b. *The probabilities p_I, p_D and p_Q become*

$$\begin{aligned}p_I &= \frac{\tau\beta_1(x)}{\psi_1(x)} + O(\frac{1}{n}) \\ p_D &= \frac{\tau}{\psi_1(x)} + O(\frac{1}{n}) \\ p_Q &= \frac{\tau^2\beta_1(x)}{\psi_1^2(x)} + O(\frac{1}{n})\end{aligned}$$

for all $x < 1$. ■

Finally, we can derive the limiting distribution for the total path. As in the case (A), we start with the limiting joint distribution of the number of I-steps, D-steps and Q-steps. We use the same notation as in Theorem 3.3. For example, we have

$$\bar{d} = 1 + p_Q \quad , \quad \alpha = \frac{1}{\bar{d}} \tag{85}$$

$$\bar{\sigma}^2 = p_Q(1 - p_Q) \quad , \quad \kappa = \frac{\bar{\sigma}^2}{\bar{d}^3}, \tag{86}$$

but in the above new values on p_I, p_D and p_Q, as estimated in Lemma 3.2b, must be used. This leads to the following results.

Theorem 3.3b. *The number of I, D and Q steps, N_I, N_D, N_Q respectively, are asymptotically Gaussian, with mean $n\mu_I$, $n\mu_D$, $n\mu_Q$ respectively, where these quantities are computed according to (53)-(55), (56) with new probabilities p_I, p_D and p_Q, as in Lemma 3.2b.* ■

Lemma 3.4b *We have $\eta_Q = \mathcal{N}(0, \tilde{\sigma}_Q^2)$ with $\tilde{\sigma}_Q^2$ given by (64) with probabilities p_I, p_D and p_Q as in Lemma 3.2b.* ■

Now, we are in position to establish our main result of this subsection, namely the limiting distribution of the total weight W_n. The next theorem proves our main finding Theorem 2.4 for case (B).

Theorem 3.5b. *The total weight W_n is asymptotically Gaussian, with mean*

$$EW_n = n\mu_W = n\sum_A m_A\mu_A \tag{87}$$

and

$$\sigma_W^2 = \frac{\mathrm{VAR}(W_n)}{n} = \sum_A \mu_A s_A^2 + \tilde{\sigma}_Q^2(m_I + m_D - m_Q)^2 \ ,$$

where the appropriate quantities must be evaluated as in Lemma 3.2b, and $\tilde{\sigma}_Q^2$ is computed from (64). ■

We concentrate now on proving Theorem 2.5, that is, estimating the total number of path $L(u)$. As discussed in Section 2, this estimate is necessary to evaluate our upper bound $\overline{\alpha}$ in Theorem 2.7.

We start the analysis with setting up a recurrence for $L(u)$. Let $f_i(j)$ be the total number of paths from O to j in i steps of the associated random walk in our grid graph $\vec{G}$. Then, $L(u) = f_n(d)$. Note that by appropriate choice of d we can analyze the number of paths for *all* cases (A)-(E). Hereafter, we concentrate on $d = un$. From Figure 4, we conclude that $f_i(j)$ satisfies the following recurrence

$$f_{i+1}(j) = f_i(j) + f_i(j-1) + f_{i-1}(j-1) \tag{88}$$

with $f_1(1) = 1$.

This recurrence is, in fact, similar to (41), and we can use the same technique to solve it. Set $g_i(z) = \sum_{j=0}^{\infty} z^j f_i(j)$ and let $\varphi(\theta, z) = \sum_{i=0}^{\infty} \theta^i g_i(z)$. After the same algebra as before, we obtain

$$\varphi(\theta, z) = \frac{\alpha_1(z)}{\theta - \theta_1(z)} + \frac{\alpha_2(z)}{\theta - \theta_2(z)} \tag{89}$$

with

$$\begin{aligned} \theta_{1,2}(z) &= \frac{1 + z \pm \sqrt{w_2(z)}}{-2z}, \\ \theta_1^* &= -(\sqrt{2} + 1), \\ \theta_2^* &= \sqrt{2} - 1, \\ w_2(z) &= 1 + z^2 + 6z \\ \alpha_2(z) &= -\frac{1}{\sqrt{w_2(z)}}, \qquad \alpha_1(z) = -\alpha_2(z) \end{aligned} \tag{90}$$

As n becomes larger, the dominant contribution comes from (89), and asymptotically we have

$$g_n(z) \sim \lambda(z) \left[\frac{1}{\theta_2(z)}\right]^n = \lambda(z)\psi_2^n(z)$$

with $\lambda(z) = -\alpha_2(z)/\theta_2(z)$.

To extract the coefficient of $g_n(z)$ we shall apply the "shift of the mean" method, as described before. We first consider only the coefficient at $g_n(z)/\lambda(z) = \theta_2^{-n}(z)$. Call it $l(u)$. Applying equation (79), as in (81), we estimate the new mean value with $\psi_1(z)$ replaced by $\psi_2(z)$ and the new $\beta_2(u)$ becomes

$$\beta_2(u) = \frac{1 + 3u^2 + u\sqrt{8(u^2 + 1)}}{1 - u^2} \tag{91}$$

and then

$$\psi_2(u) = \psi_2[\beta_2(u)] = \frac{2u\beta_2(u)}{\beta_2(u) - 1 - u(1 + \beta_2(u))} .$$

The quantity $V(u)$ is the variance related to the generating function $\frac{\psi_2(\beta_2 z)}{\psi_2(\beta_2)}$. With this in mind, it is easy to see that

$$l(u) = \frac{\psi_2(\beta_2(u))^n}{\beta_2(u)^{n(1+u)/2}\sqrt{2\pi n V(u)}}(1 + O(1/n)) = \frac{\exp(n\rho(u))}{\sqrt{2\pi n V(u)}}(1 + O(1/n)) \tag{92}$$

where $\rho(u)$ is a function of u and it is given by (21).

Now, we compute the coefficient at $g_n(z) = \lambda(z)\theta_2^{-n}(z)$, that is, we include the correction coming from $\lambda(z)$. Note that $\lambda_1(z) = \lambda(z)/\lambda(1)$ can be viewed as the generating function of a random variable. Let its probability distribution be denoted by $p_\lambda(i)$. Since the product

of two generating functions translates into the convolution of the appropriate coefficients, we have $L(u) = \sum_{i=0}^{\infty} p_\lambda(i) l(u - i/n)$. By (92) we finally obtain

$$\begin{aligned} L(u)\sqrt{2\pi n V(u)} &= \lambda(1)\sum_{i=0}^{\infty} p_\lambda(i) \exp\left(n(\rho(u) - \rho'(u)i/n + O(n^{-2}))\right)(1 + O(n^{-1})) \\ &= \lambda(e^{-\rho'(u)}) \exp(n\rho(u))(1 + O(n^{-1})) \ , \end{aligned} \tag{93}$$

where $\rho'(u)$ is the derivative of $\rho(u)$. From the above, we conclude that the constant C in Theorem 2.5 becomes

$$C = \lambda(e^{-\rho'(u)}) \tag{94}$$

where $\lambda(z)$ is given above. This completes the proof of Theorem 2.5 for case (A) and (B) (in case (A) $\rho(u)$ is given by (20)).

Remark 2. We can check Theorem 2.5 and 3.4 using the following combinatorial argument: The total number of paths T_Q with N_Q Qsteps fixed becomes

$$T_Q = \frac{(n - N_Q)!}{N_Q!\ N_I!\ N_D!} \ .$$

By Stirling's formula and tedious analysis, we can check that (64) and Theorem 2.5 lead to

$$T_Q \sim \frac{N(x) e^{-\frac{1}{2}\eta_Q^2/\tilde{\sigma}_Q^2}}{\sqrt{2\pi}\tilde{\sigma}_Q} \ ,$$

as desired. □

3.3 Case (C): $d = n - O(n^{1-\varepsilon})$

Let us first analyze a particular case, namely: $\epsilon = \frac{1}{2}$, so that $d = n(1 - \frac{x}{\sqrt{n}})$. We now set $u = 1 - \frac{x}{\sqrt{n}}$ in all formulas for probabilities and/or number of paths. For probabilities, (83) leads to

$$\begin{aligned} \beta_5(x) &= \frac{4\sqrt{n}}{x}(1 - \frac{x}{\sqrt{n}}) + O(\frac{1}{\sqrt{n}}) \\ \psi_5(x) &= \frac{2\tau\sqrt{n}}{x}(2 - \frac{x}{\sqrt{n}}) + O(\frac{1}{\sqrt{n}}) \end{aligned}$$

The below lemma is a direct consequence of the above, and the arguments used in the previous subsections.

Lemma 3.2c. *The one-step probabilities become*

$$\begin{aligned} p_I &= 1 - \frac{x}{2\sqrt{n}} + O(\frac{1}{n}) \\ p_D &= \frac{x}{4\sqrt{n}} + O(\frac{1}{n}) \\ p_Q &= \frac{x}{4\sqrt{n}} + O(\frac{1}{n}) \end{aligned}$$

with x being a given constant. ■

To prove Theorem 2.5, we proceed exactly as in section 3.2. Formula (93) holds with

$$\begin{aligned}\beta_6(x) &= \frac{4\sqrt{n}}{x}(1-\frac{x}{\sqrt{n}})+O(\frac{1}{\sqrt{n}})\\ \psi_6(x) &= \frac{2\sqrt{n}}{x}(2-\frac{x}{\sqrt{n}})+O(\frac{1}{\sqrt{n}})\end{aligned}$$

by the virtue of (91) and (84). Easy calculations also reveal that

$$\log L(x) = \rho(x)n - \frac{1}{2}\log n + O(1)$$

with

$$\rho(x) = \frac{x\log n}{4\sqrt{n}} + \frac{1}{2}(1+\log(4))\frac{x}{\sqrt{n}} - \frac{x\log x}{2\sqrt{n}} + O(\frac{1}{n}) \tag{95}$$

which establishes Theorem 2.5 in this case.

Theorems 3.4 and 3.5 are still valid. In particular, we now have $N_Q = \frac{\sqrt{n}x}{4} + n^{1/4}\hat{\eta}_Q$. This implies the next two results.

Lemma 3.4c. *The following holds*

$$\hat{\eta}_Q \sim \mathcal{N}(0,\hat{\sigma}_Q^2) + O(\frac{1}{n^{1/4}})$$

with $\hat{\sigma}_Q^2 = \frac{x}{8} + O(\frac{1}{\sqrt{n}})$. ■

Theorem 3.5c. *The total weight is normally distributed as stated in Theorem 2.3 with μ_I and μ_D computed from Lemma 3.2a, and*

$$\begin{aligned}\tilde{\sigma}_Q^2 &= \frac{x}{8\sqrt{n}} + O(\frac{1}{n}),\\ \mu_Q &= \frac{x}{8\sqrt{n}} + O(\frac{1}{n})\end{aligned} \tag{96}$$

where x is a constant. ■

We now turn to the case $d = n(1-\frac{x}{n^\epsilon}), 0<\epsilon<1$, so we set $u = 1-\frac{x}{n^\epsilon}$. This gives

$$\begin{aligned}\beta_5(x) &= \frac{4n^\epsilon}{x}(1-\frac{x}{n^\epsilon})+O(\frac{1}{n^{2\epsilon}})\\ \psi_5(x) &= \frac{2\tau n^\epsilon}{x}(2-\frac{x}{n^\epsilon})+O(\frac{1}{n^{2\epsilon}})\\ p(I,x) &= 1-\frac{x}{2n^\epsilon}+O(\frac{1}{n^{2\epsilon}})\\ p(D,x) &= \frac{x}{4n^\epsilon}+O(\frac{1}{n^{2\epsilon}})\\ p(Q,x) &= \frac{x}{4n^\epsilon}+O(\frac{1}{n^{2\epsilon}})\end{aligned}$$

$$
\begin{aligned}
\mu_I &= 1 - \frac{3x}{4n^\epsilon} + O(\frac{1}{n^{2\epsilon}}) \\
\mu_D &= \frac{1}{4}\frac{x}{n^\epsilon} + O(\frac{1}{n^{2\epsilon}}) \\
\mu_Q &= \frac{1}{4}\frac{x}{n^\epsilon} + O(\frac{1}{n^{2\epsilon}}) \\
\tilde{\sigma}_Q^2 &= \frac{x}{8n^\epsilon} + O(\frac{1}{n^{2\epsilon}}) \\
N_Q - n^{1-\epsilon}\frac{x}{4} &= n^{(1-\epsilon)/2}\hat{\eta}_Q + O(1) \qquad (97)
\end{aligned}
$$

with $\hat{\eta}_Q \sim \mathcal{N}(0, \hat{\sigma}_Q^2)$ and $\hat{\sigma}_Q^2 = \frac{x}{8}$. Theorems 3.4c and 3.5c hold with appropriate quantities computed as above.

Finally, for the number of paths we have

$$\log L(x) = \rho(x)n - \frac{1}{2}\log n + O(1) \qquad (98)$$

with $\rho(x) = \frac{x}{2n^\epsilon}(-\log x + \epsilon \log n) + \frac{x}{2n^\epsilon}(1 + \log(4)) + O(\frac{1}{n^{2\epsilon}})$.

3.4 Case (E): $s = O(1)$

Let now $r = O(1)$. Clearly, this amounts to setting $u = 1 - \frac{x}{n}$, $x = 2r, \epsilon = 1$ in all previous expansions. Furthermore, (97) becomes $N_Q = O(1)$. Also $N_D = O(1)$ and $N_I = n + O(1)$. We also have

$$\log \Pr\{Y(n) \in d\delta\} \sim n\left(\log \psi_5(u) - \frac{(1+u)}{2}\log \beta_5(u)\right) \sim -n \log \tau$$

as it should. Theorem 3.3-3.5 hold with appropriate changes.

Finally, to compute the number of paths $L(u)$ we note that (98) leads to $\log L(x) \sim \frac{x-1}{2}\log n + O(1)$. This completes the proof of Theorem 2.5.

ACKNOWLEDGEMENT

We would like to thank Professors Luc Devroye, McGill University, and Michel Talagrand, Ohio State University and Paris VI, for discussions that led to our Theorem 2.3.

References

[1] A. Apostolico and C. Guerra, The Longest Common Subsequence Problem Revisited, *Algorithmica*, 2, 315-336, 1987.

[2] A. Apostolico, M. Atallah, L. Larmore, and S. McFaddin, Efficient Parallel Algorithms for String Editing and Related Problems, *SIAM J. Comput.*, 19, 968-988, 1990.

[3] Arratia, R., Gordon, L., and Waterman, M.,rratia, R., and Waterman, M., An Extreme Value Theory for Sequence Matching, *Annals of Statistics*, 14, 971-993, 1986.

[4] Arratia, R., Gordon, L., and Waterman, M., The Erdös-Rényi Law in Distribution, for Coin Tossing and Sequence Matching, *Annals of Statistics*, 18, 539-570, 1990.

[5] R. Arratia and M. Waterman, Critical Phenomena in Sequence Matching, *Annals of Probability*, 13, 1236-1249, 1985.

[6] Arratia, R., and Waterman, M., The Erdös-Rényi Strong Law for pattern matching with a Given Proportion of Mismatches, *Annals of Probability*, 17, 1152-1169, 1989.

[7] R. Arratia and M. Waterman, A Phase Transition for the Score in Matching Random Sequences Allowing Deletions, *Annals of Applied Probability*, to appear.

[8] M. Atallah, P. Jacquet and W. Szpankowski, Pattern matching with mismatches: A probabilistic analysis and a randomized algorithm, *Proc. Combinatorial Pattern Matching*, Tucson, 1992, pp. 27-40.

[9] P. Billingsley, P., *Convergence of Probability Measures*, John Wiley & Sons, 1968

[10] J. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*, John Wiley & Sons , 1990.

[11] A. Dembo and S. Karlin, Poisson Approximations for r-Scan Processes, *Annals of Applied Probability*, 2, 329-357, 1992.

[12] Z. Galil and K. Park, An Improved Algorithm for Approximate String Matching, *SIAM J. Computing*, 19, 989-999, 1990.

[13] W. Chang and J. Lampe, Theoretical and Empirical Comparisons of Approximate String Matching Algorithms, *proc. Combinatorial Pattern Matching*, 172-181, Tuscon 1992.

[14] V. Chvatal and D. Sankoff, Longest Common Subsequence of Two Random Sequences, *J. Appl. Prob.*, 12, 306-315, 1975.

[15] J. Griggs, P. Halton, and M. Waterman, Sequence Alignments with Matched Sections, *SIAM J. Alg. Disc. Meth.*, 7, 604-608, 1986.

[16] J. Griggs, P. Halton, A. Odlyzko and M. Waterman, On the Number of Alignments of k Sequences, *Graphs and Combinatorics*, 6, 133-146, 1990.

[17] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol.I, John Wiley & Sons, 1970

[18] W. Feller *An Introduction to Probability Theory and its Applications*, Vol.II, John Wiley & Sons, 1971

[19] D.H. Greene and D.E. Knuth, *Mathematics for the Analysis of Algorithms*, Birkhauser, 1981

[20] D.L. Iglehart, Weak Convergence in Applied Probability, *Stoch. Proc. Appl.* 2, 211-241, 1974.

[21] S. Karlin and A. Dembo, Limit Distributions of Maximal Segmental Score Among Markov-Depenedent Partial Sums, *Adv. Appl. Probab.*, 24, 113-140, 1992.

[22] S. Karlin and F. Ost, Counts of Long Aligned Word Matches Among Random Letter Sequences, *Adv. Appl. Prob.*, 19, 293-351 (1987).

[23] J.F.C. Kingman, *Subadditive processes*, in Ecole d'Eté de Probabilités de Saint-Flour V-1975, Lecture Notes in Mathematics, 539, Springer-Verlag, Berlin (1976).

[24] T. Liggett *Interacting Particle Systems*, Springer-Verlag, New York 1985.

[25] G. Louchard, Random Walks, Gaussian Processes and List Structures, *Theor. Comp. Sci.*, 53, 99-124, 1987.

[26] G. Louchard, R. Schott and B. Randrianarimanna, Dynamic Algorithms in D.E. Knuth's Model : A Probabilistic Analysis, *Theor. Comp. Sci.*, 93, 201-225, 1992.

[27] C. McDiarmid, On the Method of Bounded Differences, in *Surveys in Combinatorics*, J. Siemons (Ed.), vol 141, pp. 148-188, London Mathematical Society Lecture Notes Series, Cambridge University Press, 1989.

[28] E. Myeres, An $O(ND)$ Difference Algorithm and Its Varations, *Algorithmica*, 1, 251-266, 1986.

[29] C. Newman, Chain Lengths in Certain Random Directed Graphs, *Random Structures & Algorithms*, 3, 243-254, 1992.

[30] P. Pevzner and M. Waterman, Matrix Longest Common Subsequence Problem, Duality and Hilbert Bases, *proc. Combinatorial Pattern Matching*, 77-87, Tuscon 1992.

[31] D. Sankoff and J. Kruskal (Eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley, Reading, Mass., 1983.

[32] E. Ukkonen, Finding Approximate Patterns in Strings, *J. Algorithms*, 1, 359-373, 1980.

[33] M. Waterman, L. Gordon and R. Arratia, Phase Transitions in sequence matches and nucleic acid structures, *Proc. Natl. Acad. Sci. USA*, 84, 1239-1242, 1987.

[34] M. Waterman, (Ed.) *Mathematical Methods for DNA Sequences*, CRC Press Inc., Boca Raton, (1991).

ISSN 0249-6399